Comparing visualization techniques for learning second language prosody

first results

Niebuhr, Oliver; Alm, Maria Helena; Schümchen, Nathalie ; Fischer, Kerstin

Published in:
International Journal of Learner Corpus Research

DOI:
10.1075/ijlcr.3.2.07nie

Publication date:
2017

Document version
Peer reviewed version

Citation for pulished version (APA):
Niebuhr, O., Alm, M. H., Schümchen, N., & Fischer, K. (2017). Comparing visualization techniques for learning second language prosody: first results. International Journal of Learner Corpus Research, 3(2), 250-277. DOI: 10.1075/ijlcr.3.2.07nie

Comparing Visualization Techniques for Learning Second Language Prosody – First Results

Oliver Niebuhr, SDU Innovation and Design Engineering, Mads Clausen Institute, University of Southern Denmark, Sonderborg, Denmark; olni@sdu.dk

Maria Alm, Dept. of Design and Communication (IDK), University of Southern Denmark, Sonderborg, Denmark; mhalm@sdu.dk

Nathalie Schümchen, Dept. of Design and Communication (IDK), University of Southern Denmark, Sonderborg, Denmark; nats@sdu.dk

Kerstin Fischer, Dept. of Design and Communication (IDK), University of Southern Denmark, Sonderborg, Denmark; kerstin@sdu.dk

Running Head: Prosody Visualization Techniques

Corresponding author:
Oliver Niebuhr
Ketelsenweg 6b
D-24983 Handewitt
Germany
Email: olni@sdu.dk
Phone: +49 157 74 75 333 2
Internet:
http://www.sdu.dk/en/om_sdu/institutter_centre/irca/medarbejderliste/permanent_members/oliver+niebuhr

Key words: prosody, intonation, stress, visualization, second-language acquisition, Danish, German, speech production

Abstract

We test the usability of prosody visualization techniques for second language (L2) learners. Eighteen Danish learners realized target sentences in German based on different visualization techniques. The sentence realizations were annotated by means of the phonological Kiel Intonation Model and then analyzed in terms of a) prosodic-pattern consistency and b) correctness of the prosodic patterns. In addition, the participants rated the usability of the visualization techniques. The results from the phonological analysis converge with the usability ratings in showing that iconic techniques, in particular the stylized "hat pattern" visualization, performed better than symbolic techniques, and that marking prosodic information beyond intonation can be more confusing than instructive. In discussing our findings, we also provide a portrait of the new Danish-German learner corpus we created: DANGER. It is freely available for interested researchers upon request.

1 Prosodic Notation Systems as Visualization Techniques

The speech sciences constantly gain an increasingly more detailed and comprehensive understanding of the meaningful elements of prosody and of their syntagmatic and paradigmatic structures. The growing understanding y has, together with advances in speech technology, industrial globalization and people's mobility, also increased the demands for teaching and learning of the meaningful elements of prosody across various disciplines and professions. Demands range from therapeutic measures through the integration of refugees or cross-border commuters in societies and companies to the coaching of managers and entrepreneurs in speaking charismatically.

So far it is open whether traditional visualizations of prosody are actually able to take account for all these new insights and challenges. Visualizing prosody is in fact not a new idea. Already in 1775, Joshua Steele made one of the first attempts in his "essay towards establishing the melody and measure of speech". Steele used a musical score, supplemented by vertical duration symbols. Visualizing prosody using the musical grid remained fairly common after Steele's pioneering work and was used until the middle of the 20th century, for example, by Haugen and Joos (1952), Fónagy and Magdics (1963), and Delattre (1966); see Crystal (1995) for an overview.

Driven by the works of Daniel Jones (1909) and other members of the British School of Phonetics, like Herrmann Klinghardt (1923, 1927), Armstrong & Ward (1926), and O'Connor & Arnold (1973), approaches to visualizing prosody became less specific in terms

of absolute musical intervals. Instead, they often became more specific with respect to contour slopes, syntagmatic elements of pitch and/or the marking of prominence, stress, and prosodic boundaries. Further developments of these approaches were put forward by many other scholars since then, including Pike (1945), von Essen (1956), Bolinger (1958), Delattre (1966), Isačenko & Schädlich (1970), Stock & Zacharias (1973), and Ladd (1978). More recently, several new proposals for visualizing prosodic contours have been introduced (e.g., Pierrehumbert 1980; Kohler 1997; Mehlhorn & Trouvain 2007; Prom-om & Xu 2010; Gorjian et al. 2013; Fujimori et al. 2015). Roughly speaking, the more recent a visualization technique is, the less likely it is to use the music-inspired interlinear concept that plots speech melody in between two horizontal lines representing the speaker's pitch range.

Some visualization techniques were primarily developed for teaching L2 prosody, others for exchanging information between scholars. Some are mere visualizations, whereas others represent phonological inventories of prosodic labels and symbols. Quite a few visualization techniques had originally been developed for scientific purposes and were then, without proper didactic adaptation, repurposed for teaching. Irrespective of these fundamental differences, and despite the huge diversity of visualization techniques that exist, none of these techniques has ever been formally tested with respect to its usability and suitability for untrained learners nor have there been, to the best of our knowledge, any systematic empirical comparisons of the techniques' or the learners' performances (which is not the same as measuring inter-annotator agreement, cf. Gut & Bayerl 2004).

One reason why such tests and comparisons have been neglected so far is probably that visualization techniques have predominantly been used in expert contexts. That is, experts -- very often the inventors of the techniques themselves -- used their techniques to make prosody accessible to L2 learners, research assistants (for corpus annotation), students of linguistics, or other experts (for scientific exchange and written publications).

However, new media and the growing interest in prosody across disciplines and professions require that visualization techniques are consistently applicable and, perhaps even more importantly, intuitively understandable without requiring extra effort and training of either teachers and learners. It is our experience that language teachers often find it difficult to teach intonation, either because of different priorities or because of insecurity. Visualization techniques have to be self-explanatory and compatible with self-learning situations, including digital applications, e-learning software, and computer games, as well as with frameworks in which non-experts (in the fields of prosody and intonation) teach other non-experts.

Do some already existing visualization techniques meet these requirements better than others? If so, which features are responsible for the differences in performance, and can we, on this basis, develop a new visualization technique that outperforms all existing techniques? As a first approach to answer these questions, we test the applicability and usability of six prosody visualization/notation techniques on Danish L2 learners of German without formal training in intonation. For the comparison, we have chosen six established visualization techniques for prosody representation. They differ in *which* prosodic information they represent and *how* they represent it. This concerns, for example, the degree of intonational detail (symbolic or iconic, high and low tones, straight lines, or curvy contours superimposed on or largely integrated in the text), the handling of prominence (presence or absence, gradual vs. categorical marking of stressed words), and further aspects of prosody like final lengthening and emphatic accentuation (Wightman et al. 1992; Kügler 2008; Baumann et al. 2006; Niebuhr 2010).

From the visual impressions of the compared techniques, we hypothesized that four, namely the stylized “hat pattern” visualization technique (e.g., Isačenko & Schädlich 1970), the “continuous contour” (Jones 1909), the “tadpole” visualization technique (e.g. O’Connor & Arnold 1973), and the “meandering text” (e.g., Bolinger 1958; Ladd 1978), would lend themselves well for pedagogical purposes. The remaining two notation systems, GAT 2 (Selting et al. 2009), and GToBI (e.g. Grice et al. 2005) are predominantly used in scientific contexts, yet at least ToBI is claimed by its advocates to be “intuitively” interpretable (Ladd 2008).

Although we did not know which technique(s) would turn out as most suitable for teaching, it is reasonable to assume that useful visualization techniques somehow manage to strike a balance between oversimplifying on the one hand and overwhelming their users with too many details on the other. This logic predicts that the stylized “hat pattern,” for example, can be expected to perform better than the “continuous contour”. The former is directly integrated in the text and refrains from showing too much intonational detail, particularly detail that is not phonologically relevant and anyway produced by speakers due to aerodynamic or biomechanical reasons (cf. micro-prosody, Kohler 1990). Further assumptions are difficult to make. For example, the “tadpole” technique could either perform worse than the "hat pattern" technique as "tadpoles" also display complex and not only phonologically relevant intonation details; or "tadpoles" could perform better than the “hat pattern” technique as they include iconic prominence marking.

All six systems were tested and compared with respect to (a) consistency, i.e. if the notation systems make all speakers produce similar intonation contours, and (b) correctness, i.e. the learners' success in producing native-speaker like prosodies. We were also interested to see how easy or difficult our participants thought the tested visualization techniques were.

2 Method

In order to test the usability of established visualization techniques for prosodic information in second language teaching contexts, a production study was conducted. We decided to use German as the language of study because German prosody is well researched, the team includes an expert on German intonation, and our local university programs include teaching of German as a second language.

2.1 Participants

We recorded 20 participants who worked with the visualization techniques in pairs. One data set was lost due to a corrupted memory card. Thus, our analysis only included the remaining nine recorded speakers pairs, i.e. 18 participants in total: 14 females, 3 males, 1 non-binary (genderqueer). The participants' age ranged from 16 to 59 years, with 72 % being between 20 and 26 years old. All participants were Danish citizens and spoke Danish as their native language. Furthermore, all reported to speak German on a fluent (communicative) level, and most (83 %) stated the same about English. German was generally obtained in German-as-a-second-language courses either in the form of A-level high school courses or as part of the participants' study program. While German pronunciation was reportedly part of these courses, participants had not received any training on intonation or other prosodic features. At the time of the study, all participants were either in high school or university students at bachelor, master, or PhD level.

2.2 The Stimuli

2.2.1 Sentence Material

The stimuli were selected from the so-called Berlin Sentences (German: *Berliner Sätze*). This is a corpus of 102 German statement and question sentences that consist of 3-7 common

words. The sentences were designed to be balanced and representative with respect to segmental sound patterns. This includes that the sentences roughly reflect the relative frequencies with which the individual German phonemes occur in everyday language use. Moreover, the sentences intended to cover all combinations of two sounds (i.e. all diphones) that are known to occur in German (Sotscheck 1984). The sentences differed in length (7-14 syllables) but were created in such a way that readers can produce them as a single prosodic phrase. The latter fact, which found empirical support in the study of Peters (1999) and also proved to be true in the present study, was the main reason for us to choose this speech material for our study.

We assumed that the criteria of phonological and phonetic balance that guided the construction of the Berlin Sentences would give the sentences a constant moderate level of difficulty concerning their pronunciation by non-native speakers of German with the same Danish L1 background. The feedback that we got after the experiment as well as our own observations during the experiment were in line with this assumption. Moreover, all our participants were familiar with German pronunciation and had time to practice each sentence until they were happy with their production.

Furthermore, the Berlin Sentences consist of common words, which made them easy to understand by our advanced L2 learners of German. A final reason for using the Berlin Sentence was that they are an established resource in phonetic studies on both sound segments (Feldes & Herzog 1997; Johner et al. 2012) and prosody or intonation (Möbius 1993; Peters 1999; Wahlster 2000).

For each of our six visualization techniques we selected two sentences, one statement and one question. The selected target sentences were prosodically comparable. Each sentence elicits, in a native-speaker's rendering, a sequence of prenuclear and nuclear pitch accent. The two pitch accents are separated by at least two unaccented syllables in order to avoid production and analysis artifacts due to tonal crowding (Caspers & van Heuven 1993). For the same reason, only sentences with at least five words were chosen. Furthermore, we favored those sentences with as many voiced sounds as possible in order to optimize conditions for the evaluation of the participants' intonation contours after the experiment.

Our Berlin Sentences were not just designed to be balanced in terms of pronounceability and comprehensibility.They were also discussed and practiced by our participants prior to producing the final version for our analysis. For these reasons, we considered 'sentence' a

negligible factor for the results of the present study. Considering 'sentence' a negligible factor means also that we do not distinguish between different types of questions, in particular, between wh-questions, deliberative *if*-questions, and verb-initial yes-no questions. This seems justified since, firstly, all three types of questions *can* basically be realized with a final rising intonation in German. Secondly, our L2 speakers never lived in Germany and thus could not know that the three question types are differently likely to occur with final rising and falling intonations in German spontaneous speech (cf. Niebuhr 2015). On the contrary, textbooks for L2 learners of German still state that *all* questions are to be realized with rising intonation, cf. the introductory comment in Batliner (1991: 147). For instance, Griesbach (2000:231) advises L2 learners of German to identify questions by means of question marks or rising intonations at the ends of written or spoken sentences. The textbook of Harst et al. (2015) instructs language teachers to explain (with supporting hand movements) to L2 learners of German that German sentence melodies rise in questions and fall in statements. Thirdly, and most importantly, our speakers were explicitly instructed to produce the question sentences with the intonation contour indicated by the given visualization, which was in all cases the stereotypical final rise. Accordingly, data inspections showed no intonation biases due to question type.

2.2.2 Prosodic reference productions

Our twelve question and statement sentences were produced and recorded with clearly pronounced prosody by a native speaker of (Northern Standard) German. Statements ended in a terminal falling and questions in a high rising intonation. As was stated in 2.2.1., each sentence was produced with a prenuclear and nuclear pitch accent, separated by at least two syllables. The clear native-speaker prosodies were used as reference productions in our study. That is, the reference productions determined how the sentences were represented in terms of their corresponding visualization/notation technique.

For the orthographic representation of the sentences, we refrained from using any punctuation marks in order not to unnecessarily confuse the participants, make them focus on the visualization alone and avoid providing them with additional or even conflicting prosodic cues.

2.2.3 Visualization and notation techniques

The “hat pattern” visualization technique (e.g., Isačenko & Schädlich 1970) consists of stylized intonation contours that are largely integrated in the text. The “hat pattern” focuses on the tonal turning points while disregarding smaller tonal movements between the turning points, as can be seen in Figure (1). In this respect, the “hat pattern” technique follows the idea of the IPO model (‘t Hart et al. 1990). Note that we will use the term “hat pattern” here in a conceptual sense that also applies to our question intonations in which the “hat” is actually turned upside down both phonetically and phonologically.

Figure (1): The stylized “hat pattern” visualization technique, drawn according to the reference productions.

Similar to the “hat pattern”, but more closely mirroring actual intonational details, is the visualization technique developed by Jones (1909). We will refer to this technique as the “continuous contour” technique. However, we adopted it from Jones without the musical grid background, in order to prevent our participants from using a sing-song intonation. The "continuous contour" technique maintains the natural "ripples" (i.e. smaller rises and falls) between the main tonal turning points. In comparison to computer-generated F0 contours, the manually drawn “continuous contour” shows no micro-prosodic F0 movements and is not interrupted by unvoiced sounds (cf. Kohler 1990). Real F0 contours would not have fit into our test sample, and they would required knowledge of physical speech signals, their interpretation and manual correction (e.g., of octave errors), which runs counter to the intuitively applicable visualizations we are looking for.

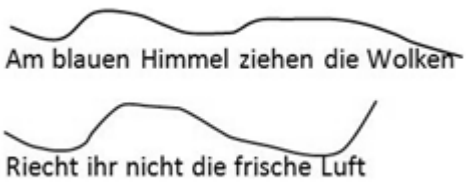

Figure (2): The “continuous contour” visualization technique, drawn according to the reference productions.

The “tadpole” visualization contour, as shown in Figure (3), originates from the British School of Intonation, see O’Connor & Arnold (1973) for major representatives of this school and Klinghardt (1923, 1927) for earlier uses of a similar visualization technique. It consists of dots representing the relative height of the syllable. The dots vary in size and in this way not only indicate stressed syllables but also reflect, in a gradual fashion, the syllables' perceptual salience levels. Moreover, important tonal movements are indicated by turning the dot into a “tadpole”, the tail of which indicates direction and shape of the tonal movement. Slightly refining this concept, we separated the pre-nuclear and nuclear accent dots, which are the closest to each other in size, by using a hollow dot for the pre-nuclear accent.

Figure (3): The “tadpole” visualization technique, drawn according to the reference productions.

Another visualization technique, which goes back to Bolinger (1958) and was later adopted, for example, by Ladd (1978) and Fagyal (1997), turns the text itself into an intonation contour. We will refer to this visualization technique as “meandering text”, see Figure (4).

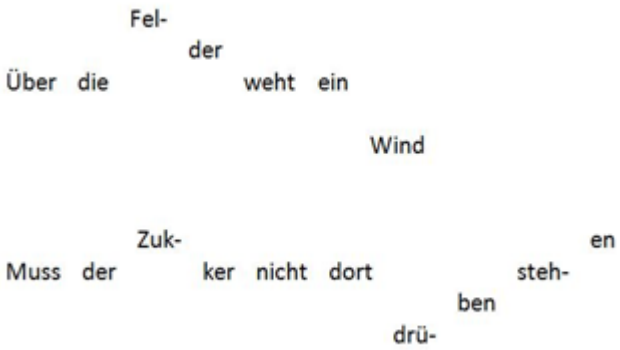

Figure (4): The “meandering text” visualization technique, designed according to the reference productions.

Rooted in conversational analytic and interactional linguistic traditions, the GAT 2 system relies heavily on Gail Jefferson's transcription conventions with additional symbols to represent prosody in talk-in-interaction (Selting et al. 2009). For practical reasons, the notation technique only makes use of signs and symbols available on a standard keyboard. Stressed syllables are marked with capital letters. The tonal movements on the stressed syllables are marked by accent signs before the onsets of those syllables. For instance, the acute accent <´> means a rising pitch movement; the grave accent <`> represents a falling pitch movement. Prolonged sounds are marked by colons, as, for example, in the final lengthening of *WE:G* (Engl. "ROA:D"), see Figure (5). Arrows are used to mark upward or downward pitch "jumps".

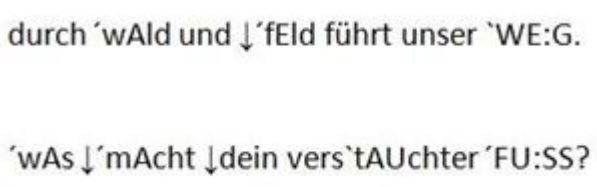

Figure (5): The GAT 2 notation system for the transcription of spoken German. The used symbols represent the reference productions.

GToBI ("German Tones and Break Indices", e.g., Grice et al. 2005) is the most established German version of the American transcription system ToBI (Pierrehumbert 1980; Beckman & Pierrehumbert 1986). The German ToBI system was developed to facilitate the exchange of prosodically annotated data, and to describe German intonation patterns in as much detail as is phonologically required (see "Übungsmaterialien" http://www.gtobi.uni-koeln.de/ku_gtobi.html). Especially important for the annotation is the percent symbol <%> which marks the so-called boundary tones at the end of the intonation phrase (e.g., L%), and, optionally in the case of a low tone, also at the beginning of the intonation phrase (%L). The asterisks <*> show the most salient tonal event, i.e. the relative pitcg target an accented syllable is associated with. These most salient tones can be combined with leading and/or trailing tones, in this way creating tonal movements, such as rises (L+H*, L*+H) or falls (H*+!H*, H+L*).

In the GToBI condition, we indicated to our participants that the symbols <H> and <L> stand for “high” and “low”, see Figure (6). However, we did not explain the other symbols surrounding these letters as this would have countedacted our aim to test the systems’ intuitiveness.

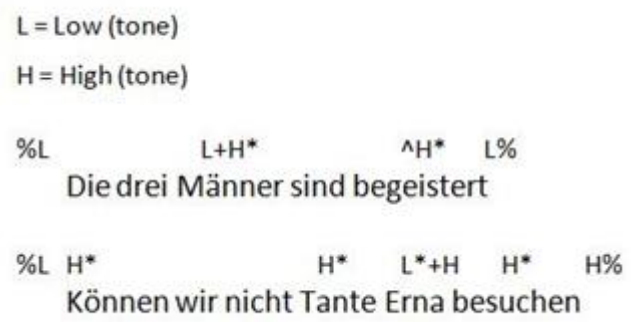

Figure (6): The GToBI notation system for German prosody. The used symbols represent the reference productions.

2.3 Procedure and data elicitation

In the production study, we were not only interested in the participants' performances with the different visualization techniques. We also wanted to know what techniques they *thought* were the easiest to use and *why*. Consequently, we had the participants working in pairs, employing a variant of the think-aloud method (Boren & Ramey 2000), which engages participants in dialogue.

In order not to influence the participants’ evaluations of the prosody visualization techniques, for instance, by using ordered or hierarchically structured reference frameworks like numbers from (1) to (6) or letters from (A) to (F), we referred to each technique by a color. The color was shown and stated in the heading and in the evaluation scale following each visualization. Each visualization technique was printed on a separate sheet of paper and handed out to the participants in random order. An example of a sheet is depicted in Figure (7). It contains a colored heading in the same color as the name, some bullet-points briefly repeating the task instructions, and an evaluation scale in the same color as the heading.

German original:

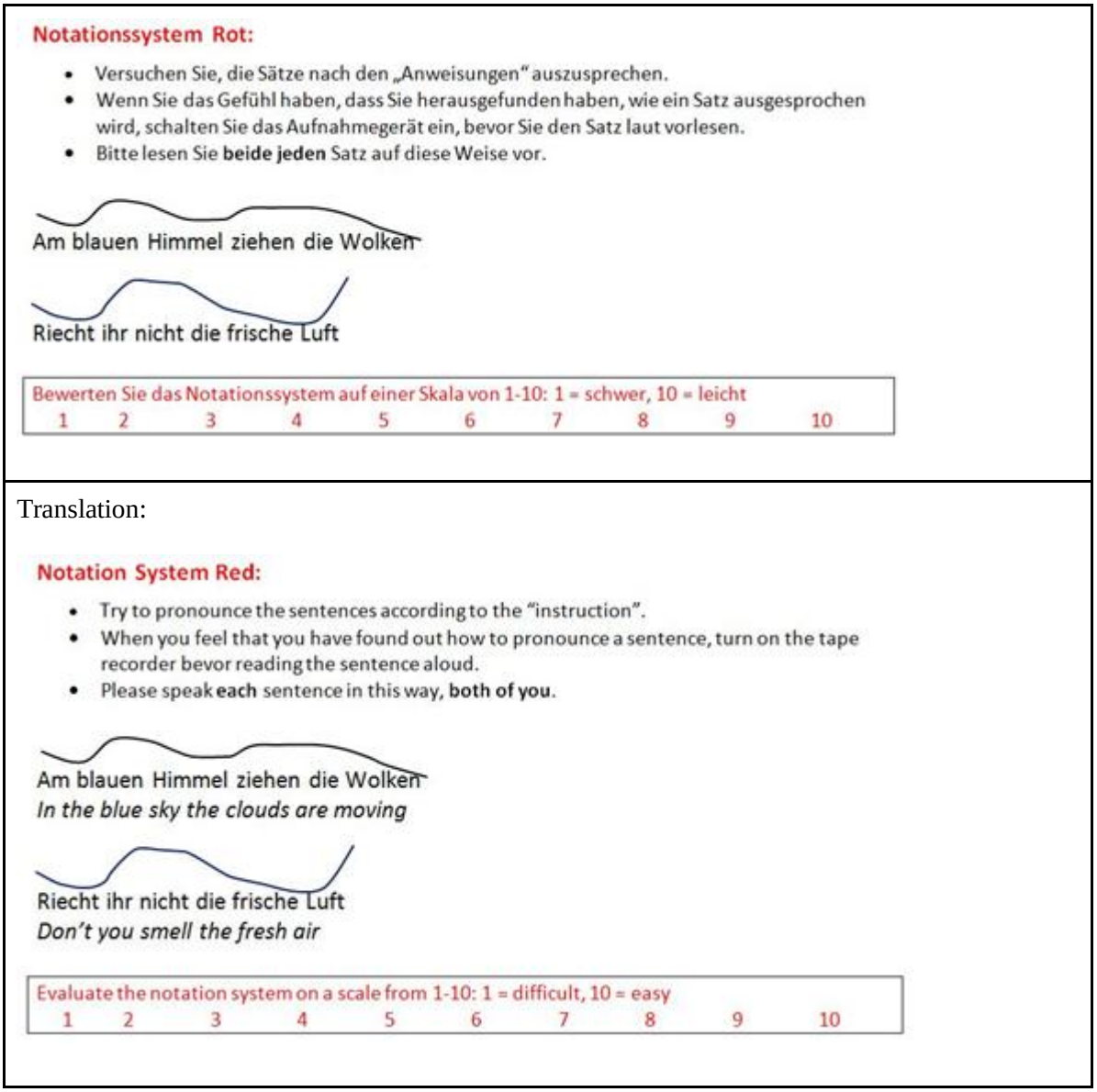

Figure (7): Example of a working sheet for presenting the visualization techniques.

The participants were instructed to choose any visualization to begin with and discuss how the sentences should be realized. Once they had settled on a realization, they were asked to produce the corresponding sentence in one go.

As is shown in Figure (8), the participants were seated side-by-side at the table in a quiet room with the visualizations in front of them. A stationary camera (model GoPro Hero) was placed about 1.5 meters away from the participants. The camera filmed the entire session and also recorded the audio signal in regular CD quality, i.e. 44.1 kHz sampling rate and 16-bit quantization. In addition, we had a separate digital voice recorder lying on the table as a

back-up. The present study used the camera signals, as their quality was sufficient for an auditory analysis.

Figure (8): Images of three of our speaker pairs, illustrating the setup of the recording session and the procedure speech-data elicitation.

Following the production of the sentences, the participants were asked to evaluate each visualization technique on a scale from 1 (difficult) to 10 (easy). Finally, they were asked to rank all techniques from easiest to most difficult on a separate sheet, providing brief explanations for their decisions. The participants needed between 20 and 40 minutes to complete the task.

2.4 Data Analysis

2.4.1 Prosodic-phonological annotation

The prosodic-phonological analysis was conducted on the basis of the Kiel Intonation Model (KIM, Kohler 1997, Niebuhr 2013). The KIM is a finely differentiated phonological intonation model. It is applied on an auditory basis and rooted in the tradition of contour models. That is, the basic building blocks of intonation are assumed to be rises, falls, and combinations thereof (peaks and valleys) rather than sequences of high(er) and low(er) tonal targets. A further characteristic of the model is that it considers the temporal alignment of the intonational building blocks with the segmental string to be directly phonologically relevant

rather than trying to derive a rule-based alignment from an abstract underlying tune-text association. It is also one of the very few intonation models that distinguishes between different phonological levels of perceptual salience of syllables and allows all levels (except for level 0, see Baumann et al. 2016) to be combined with each type of pitch accent. This, in combination with the large sets of paradigmatic phonological contrasts, makes the KIM a very handy, practice-oriented model, especially when it comes to complex patterns like spontaneous speech and L2 productions. Recent developments of the KIM are included in the new DIMA initiative whose aim is to merge the key models of German intonation into a single consensus model (Kügler et al. 2015).

The participants' individual target-sentence realizations were annotated in terms of their meaningful phonological KIM categories. The annotation was based on the symbol inventory of KIM, PROLAB (Prosodic Labeling), see Peters & Kohler (2004). The annotation consisted of two consecutive steps: (1) identifying pitch-accented words and giving them either weak, normal, or emphatic prominence levels (cf. Baumann et al. 2016); (2) determining the melodic categories that were associated with the stressed words and with the melodic boundaries in between the words and to both sides of the sentences.

Steps (1) and (2) were both carried out on an auditory basis by the same trained expert, who has more than 10 years of experience in annotating prosodic data (the 1st author). With reference to this experience, it is important to point out that prosody perception, even through a trained ear, is never perfectly precise, and assigning prosodic labels always leaves some room for interpretation. Consequently, some fuzziness and subjectivity are inevitable in our auditory-analysis procedure. However, this does not mean that the results obtained by this procedure are not reproducible and valid.

First, studies like Gut & Bayerl (2004) have shown that experience with annotations positively influences reproducibility, especially for complex phenomena like prosody. Second, it is reasonable to assume -- and supported by empirical evidence -- that intra-annotator agreement can be higher than inter-annotator agreement. Thus, having all annotations done by a single experienced expert should have given us the best possible outcome for the applied procedure -- an outcome that is not more variable and subjective than an annotation conducted by multiple annotators. Third, if the annotations contain a certain degree of intra-annotator variability, then this variability is equally (and randomly) distributed across all target sentences and types of prosodic phenomena and their

paradigmatic contrasts. That is, it is highly unlikely that our annotation procedure created artificial differences in the performances of our six compared visualization techniques.

With regard to the phrase-final rises and falls, it is probably even an advantage to address questions like those of the present study in the first step with analyses based on auditory labels rather than acoustic measurements. For example, unlike for English and German (Grabe 1998; Niebuhr & Dilley 2016), the truncation of final F0 movements has never been investigated for native speakers of Danish. It cannot be ruled out that Danish speakers truncate final F0 movements and apply this process also in L2 German. None of our target sentences represents an extreme truncation condition. Nevertheless, our speech material could still include various degrees of truncation of final F0 movements. Truncation is primarily an acoustic phenomenon that is compensated for in speech perception (see Rathcke 2013; Niebuhr & Dilley 2016). Thus, our auditory PROLAB labeles could actually represent the sentence-final intonations more consistently than the underlying F0 movements with their various degrees of truncation.

Regardless of all arguments in favor of an auditory analysis, it is of course planned, as a follow-up study, to complement the prosodic annotations by objective acoustic-prosodic measurements. These measurements are currently conducted and will be presented in a separate paper. However, in order to further dispel doubts about our present data, we can already state that phonological differences between annotation categories are consistently paralleled by significant differences between acoustic measurements. This fact supports that the annotation categories of PROLAB were systematically applied to our target sentences.

As is explained in Kohler (1997) and Niebuhr (2013), the KIM includes five pitch-accent categories, three rising-falling peaks and two (falling-)rising valleys. The peak and valley categories are distinguished by their timing relative to the accented-vowel onset. In addition, the KIM has five different phrase-final, two different phrase-initial, and three different concatenation contours (between two pitch accents), each of which is phonologically defined by its pitch scaling relative to the speaker's pitch range or adjacent pitch-accent peak or valley levels. Thus, for a prosodic phrase with two fully-fledged normally prominent pitch accents, the KIM would basically allow distinguishing 750 different intonation patterns. For three pitch accents, the number increases further to 11,250 different intonation patterns, and this does not even include weak and emphatic prominence levels. Of course, many of these phonological prominence and intonation sequences are either not possible phonetically, or they do not occur in German for functional reasons. Yet, these numbers give an idea of the

value and level of detail of our contrastive evaluation of the performance of the six prosody visualization techniques.

2.4.2 Contrastive evaluation and statistics

The contrastive evaluation of the six prosody visualization techniques took place in two ways: A first analysis compared the *consistencies* of the visualization techniques. To that end, the number of phonologically equivalent realizations (= agreeing annotations) was determined for each word of a sentence across all participants. These agreements were summed up for each sentence and set in relation to the maximum possible number of agreements (= all subjects realized every word prosodically in the same way). The resulting percentage indicates how well each visualization technique performed in eliciting the same sequences and combinations of prosodic categories from the participants.

The second contrastive evaluation concerns the *correctness* of the prosodic patterns elicited with the different visualization techniques. For this, we counted, again on a word basis, in how many instances the participants' productions agreed in terms of the phonological KIM categories with the corresponding native-speaker reference production. These agreements were added up for the entire sentence and set in relation to the maximum possible number of agreements (= all participants realized every word in the same *and* prosodically correct way). The resulting percentage indicates how well each visualization technique performed in eliciting the prosodic patterns that the native German reference speaker produced (note that 'correct' only means agreement with the native speaker's prosody and not that are the visualizations show the only viable and allowed prosodic realizations of the sentences).

The KIM allowed us to differentiate between three types of errors in this correctness analysis. This first type of error is a lack of agreement in the pitch-accent category. It is referred to as *pitch-timing error*, taking into account that it is the peak and valley timing relative to the accented vowel that is considered phonological in the KIM. The second type of error is a lack of agreement in the phrase-final/-initial or concatenation category, referred to as *pitch-scaling error*. We used the term 'scaling' here as all errors of this type result from F0 landmarks whose placement along the frequency axis is wrong relative to that of the adjacent pitch-accent peak or valley, thus creating, for example, plateaux instead indentations or final rises instead of final falls. The third type of error represents a lack of agreement in the prominence level of a pitch accent, referred to as *stress-level error*. That is, stress in the present study is a

post-lexical phenomenon rather than a binary feature of certain syllables within a word (following the traditional definition of stress, see Abercrombie 1991; Ladd 2008).

Chi-squared tests, based on the absolute frequencies behind our percentages, were conducted in order to test for significant differences in the performance of the six visualization techniques both between and across questions and statements.

3 Results

The analysis shows that some notation systems produce more errors than others, and that, in general, questions are more problematic (i.e. error-prone) than statements ($\chi^2[5]=38.7$, $p<0.0001$). As Figure (9) indicates, the "hat pattern" visualization technique produces the least errors regarding both questions and statements. The "continuous contour" and the "tadpole" notations both perform almost equally well, but overall worse than the "hat pattern". The total number of errors triggered by the other three visualization techniques, i.e. the "meandering text" technique, GToBI, and GAT 2, is between 20 % and 100 % higher.

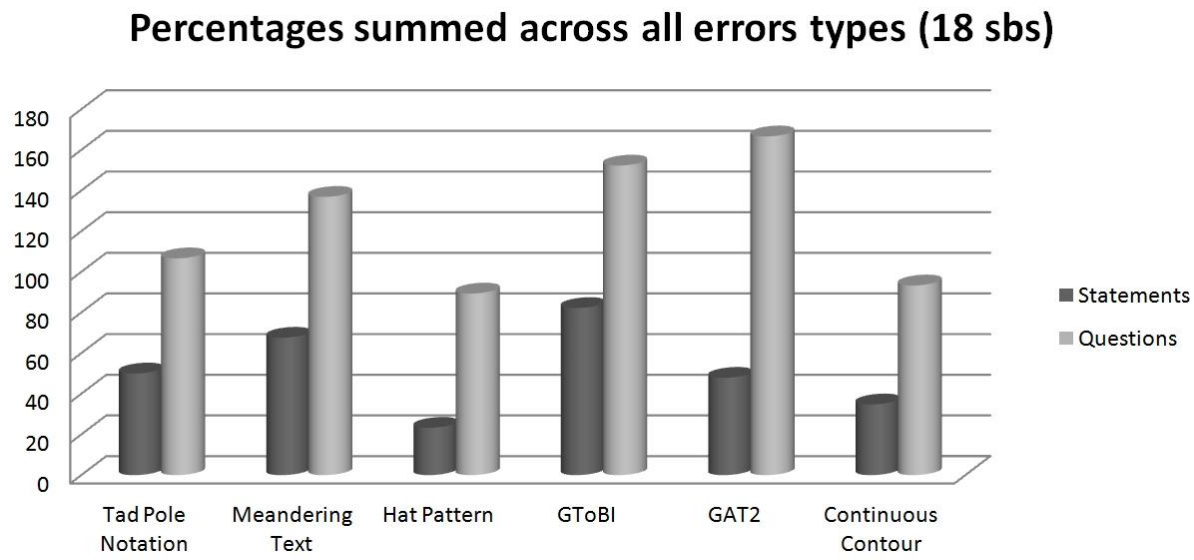

Figure (9): Percentages of errors in the correctness analysis, summed across the 18 subjects (sbs) and all three error types (sums can therefore be higher than 100 %).

Furthermore, the types of errors participants made are distributed differently across the six visualization techniques. One major example is the pitch-timing errors in statements

($\chi^2[5]=34.8$, $p<0.0001$), see Figure (10). While "hat pattern" and "continuous contour" only caused a negligible number of errors in the timing of pitch accents (3-10 %), the "meandering text" and GToBI made participants produce almost every second pitch accent with the wrong timing (i.e. 40-50 % of phonologically incorrect pitch-accent categories). A similar picture emerges for pitch-scaling errors ($\chi^2[5]=26.3$, $p<0.001$). Those participants who produced the target sentences' prosodies based on the "hat pattern" made no mistakes at all (0 %). The "tadpole" notation triggered only about 5 % pitch-scaling errors in statements. The error rates caused by the "meandering text" and GAT 2 are 4-5 times higher.

For questions (see Figure (11)), both pitch-timing and pitch-scaling errors are, as previously mentioned, overall considerably and significantly higher than for statements, but not differently distributed across the six visualization methods. What we found most striking is that most scaling errors in the case of questions are actually due to the fact that participants realized questions with a *falling* rather than a rising intonation, even in the presence of clearly contradicting visual cues from iconic visualization techniques like "hat pattern" and "continuous contour". Significant visualization-specific error rates in questions only concerned stress level ($\chi^2[5]=12.0$, $p<0.05$). GToBI and GAT 2 performed worst in this respect, whereas the "hat pattern" was again among the top performers. Note that those techniques that explicitly marked stress levels and/or stress positions, such as the "tadpole" visualization, did *not* produce less but *more* stress-level errors than the other techniques, including the "hat pattern" that marked neither stress level nor stress position. In general, stress-level errors were less frequent than both pitch-timing errors ($\chi^2[5]=31.0$, $p<0.0001$) and pitch-scaling errors ($\chi^2[5]=37.5$, $p<0.0001$).

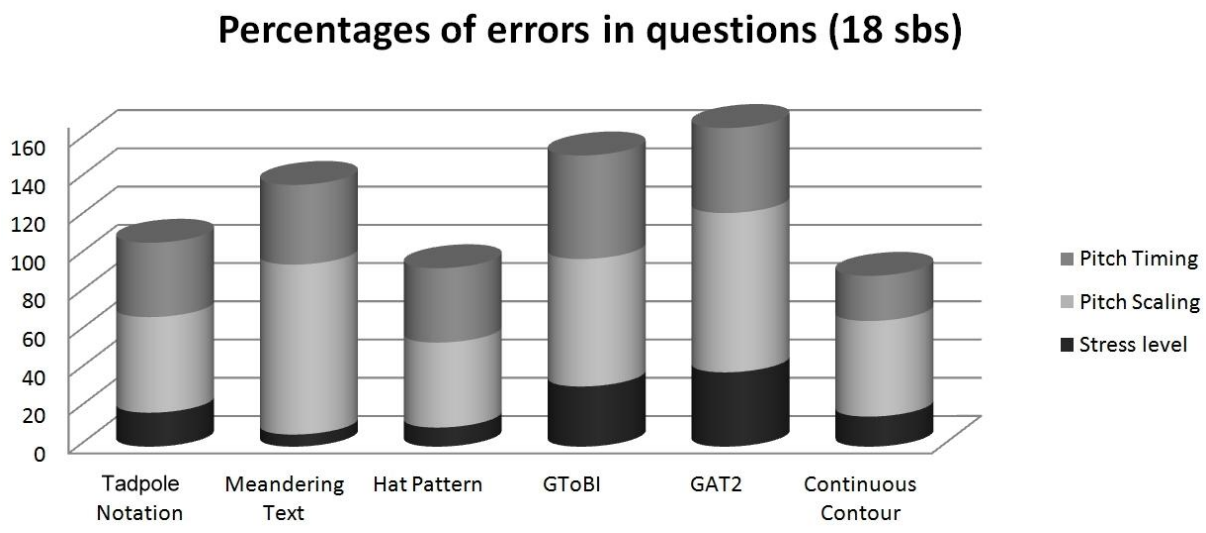

Figure (10): Percentages of errors across all 18 subjects (sbs) in the correctness analysis of statements. Percentages are displayed separately for the three different error types (sums across the three error types can be higher than 100 %).

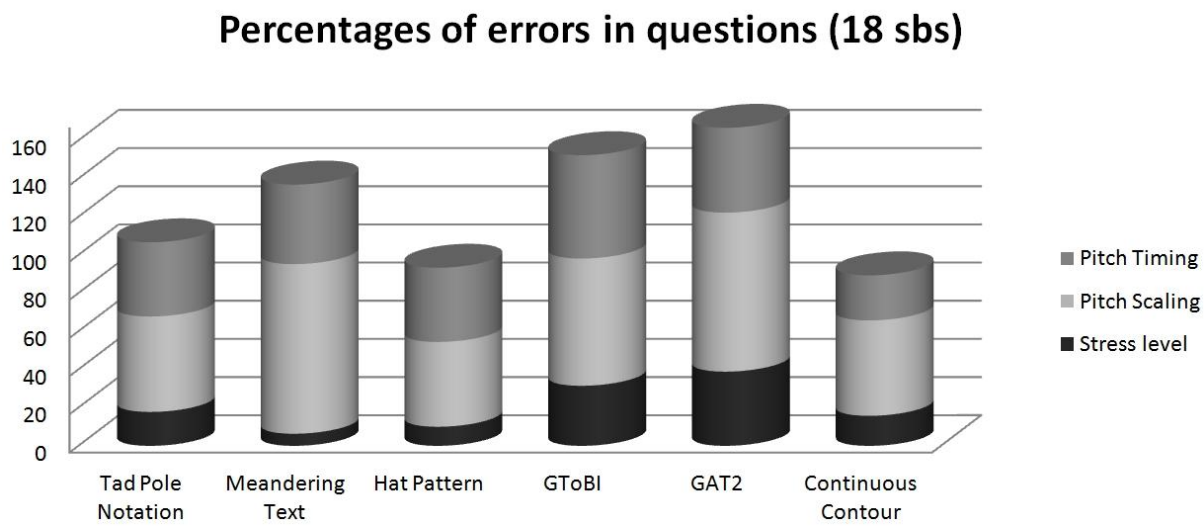

Figure (11): Percentages of errors across all 18 subjects (sbs) in the correctness analysis of questions. Percentages are displayed separately for the three different error types (sums across the three error types can be higher than 100 %).

Our results concerning the consistency of the participants' productions are summarized in Figure (12). First, it shows that consistency is overall fairly high. This means that all visualization techniques were capable of making a majority of speakers do "the same thing" (whether correct or not). Only GToBI and GAT 2 almost fall below the 50 % majority threshold for consistency, particularly in the case of questions. Second, questions were realized 10-30 % less consistently than statements, independently of which visualization technique supported the participants' realizations ($\chi^2[5]=16.3$, $p<0.01$). Third, when both questions and statements are taken into account, the "hat pattern" elicited the most consistent prosodic realizations across the 18 subjects, followed by the "meandering text" and the "tadpole" visualization ($\chi^2[5]=13.9$, $p<0.05$).

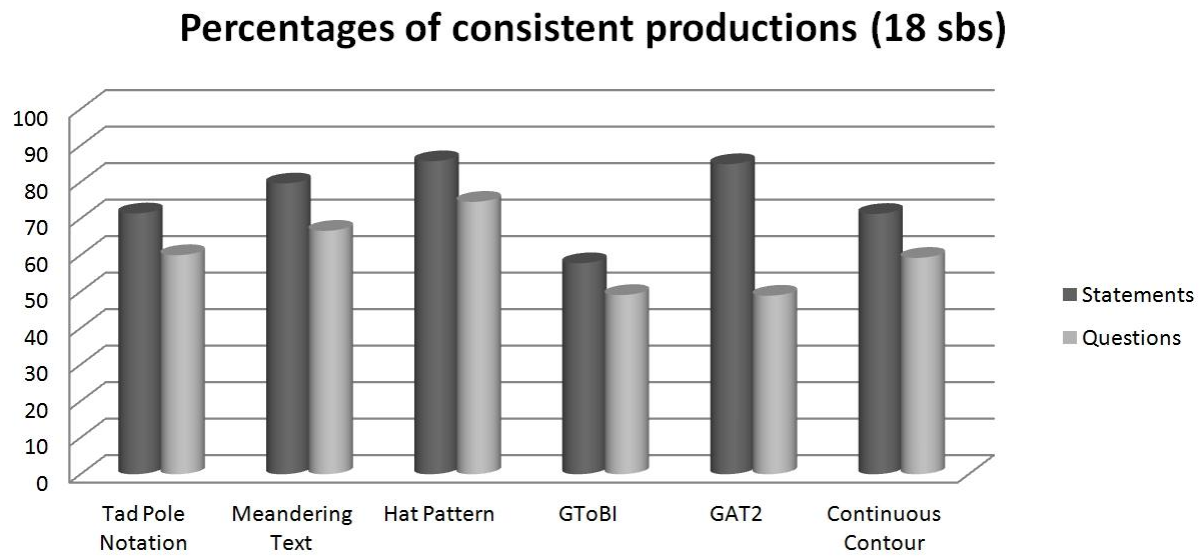

Figure (12): Percentages of consistent (but not necessarily correct) statement and question productions elicited by the six visualization techniques across all 18 subjects (sbs).

4 Discussion and Conclusions

We begin our discussion by linking the results of the auditory-phonological analysis with the analysis of the participants' feedback regarding the usability of the visualizations techniques. Participants' feedback was summarized from the think-aloud protocols that were filled out by a speaker pair each time they were finished with a visualization technique.

The "hat pattern" was the visualization technique that showed the best results regarding consistency and correctness in the auditory-phonological analysis. It was also ranked highest by the participants. According to the participants, the "hat pattern" was also easy to understand and apply, and it gave a good overview of the entire progression of the sentence. The "meandering text" technique was ranked second by the participants. It was described as straightforward, intuitive, and easy to understand, not least through its clear visual design. However, the visual design of the "meandering text" was judged to be initially challenging, and some participants missed a clear description of what the peaks and valleys formed by the words actually stood for: "Did they indicate pitch or volume?" In the auditory-phonological analysis, the "meandering text" technique only yielded the third-best results. Furthermore, participants critically commented on the 'chunkiness' of the visual presentation in the "meandering text" technique.

Ranking third on the participants' side, the "continuous contour" was considered understandable without further explanation, but at the same time – especially in comparison with the visually similar "hat pattern" – unclear and hard to apply due to its continuous, smooth and thus relatively flat pitch shape. It is probably this lack of edges and angles and the resulting unclear tune-to-text association that made the "continuous contour" perform worst among the iconic systems in terms of stress-level errors and overall consistency. Only the symbolic systems GAT2 and GToBI generated more stress-level errors and were worse in overall consistency than the "continuous contour".

The "tadpole" notation was ranked fourth by the participants. The participants' comments included positive feedback (indication of stress, easy to follow, precise) as well as negative comments (not intuitive to understand, meaning of symbols unclear, difficult to apply). The auditory-phonological analysis showed that the "tadpole" visualization technique primarily produced errors with respect to stress level. This is particularly striking as the "tadpoles" provided the most detailed visualization of stress placement and level. The high error rate can only mean that this level of detail was more confusing than instructive.

Finally, GAT 2 and GToBI scored lowest in both the participants' ranking and the auditory-phonological analysis. GAT 2 was mostly described as confusing and unclear concerning the meaning of the capitalized letters and (punctuation) symbols. Furthermore, the participants stated that the visualization technique was hard to apply and, due to its complexity, stimulated confusion and over-analyzing. However, some participants also stated that a brief explanation would probably improve understanding and application considerably. GToBI did not receive any positive comments at all. Participants were confused by the largely opaque meaning of the symbols, their placement and even the interpretation and implementation of "high" and "low" in general. In agreement with this usability assessment, GToBI was the worst system in terms of production consistency. It also triggered the most prosodic errors in statements and the second most prosodic errors – after GAT 2 – in questions, with pitch-accent timing being the most frequent error type. This is somewhat surprising given that, unlike GAT 2, GToBI explicitly specifies and shows sequences of H and L and distinguishes them from single H and L tones. This can only mean that the participants in fact did not understand the meaning of "high" and "low" and were moreover unable to interpret the asterisks <*> after high and low tones.

Note in this context that we do not want to raise doubts about the scientific value and usefulness of symbolic notations like GAT 2 and GToBI. We were concerned here with the

intuitiveness of these notations when they are used for eliciting prosody by naive speakers rather than for annotating and analyzing prosody by trained users or researchers. These are two completely different fields of application, and our results are therefore not incompatible with the fact that, for example, ToBI-based systems show a high inter-annotator agreement when applied by trained users (Breen et al. 2012; Kügler et al. 2015).

In summary, the following conclusions can be drawn from our results, bearing in mind that we used only Danish participants and hence have to leave open the question of cross-language generalization. There are considerable differences between existing visualization techniques regarding their ability to intuitively elicit consistent and correct sentence prosodies from second-language learners. The most pronounced differences were found between the group of symbolic techniques on the one hand and the group of iconic techniques on the other. In line with our hypotheses, our data shows that symbolic visualization techniques like GToBI and GAT2 were less intuitive for teaching L2 prosody than iconic visualization techniques.

Furthermore, those techniques that included other prosodic information besides the intonation contour, such as stress levels and sound prolongations, caused *more* rather than fewer errors. This strongly suggests that too much prosodic detail, even if it is phonologically relevant, is confusing and causes overload to the students. If we assume a natural parallelism between changes in the intonation contour on the one hand and increases in duration and acoustic energy (i.e. syllable prominence) on the other (cf. Kohler 1991), then marking prominence in addition to the intonation contour could indeed be redundant, at least for Germanic languages. Remember in this context that stress level is not a binary feature of certain syllables within a word. The term refers here in a post-lexical way to the absence and presence of pitch accents and, in the latter case, their level of perceptual salience. Thus, it is not possible that the relatively low error rate for stress level (as compared to timing and scaling errors) is caused by speakers' knowledge of the (L2) lexicon. Rather, it could be due the fact that content words, and nouns in particular, attract pitch accents whereas function words are typically realized non-accented. As this is the same in Danish and many other languages, it inherently reduces the range of possible stress-level errors. Nevertheless, adding stress-level markings could be useful for visualization techniques; and one way of avoiding a potential information overload in prosody visualization could be to present intonational and stress-level information subsequently or in separate visual displays. There may also be

interpersonal differences in the preference for certain visualization techniques – an issue that will become analyzable with increasing data elicitation.

As expected, the “hat pattern” is easiest to understand and leads to the least problems and highest consistency. On the whole, none of the visualization techniques analyzed provides a fully transparent and fail-safe method for students to recover the intended intonation contour of a sentence, which limits their use as a teaching tool. Thus, another conclusion from our production study is that applying visualizations of prosodic information to novel sentences is, in general, not at all straightforward. It is just as Anderson (2004:92) states in view of his own pedagogical experience: "Simply providing the student with text with interlinear pitch markings in the form of dots or lines does not work well."

The "meandering text" involves a similar intonational stylization as the "hat pattern", but it is the text itself rather than a separate contour embedded in the text that represents the stylized intonation. The "tadpole" visualization is intonationally as specific and superimposed on the text as the "continuous contour" visualization. The only differences are that the former technique includes stress-level marking and that the latter contour is not interrupted. Thus, even what at first glance appears to be a small visualization difference can have a large effect on the correctness and/or consistency of prosody elicitation. This is an important insight for both prosody research and teaching and shows the need for further more detailed investigations and comparisons of visualization/notation techniques. The distinguishing characteristics of the “hat pattern”, in particular its stylized, edgy representation of the contour and its spatial proximity to the text, could possibly constitute a starting point for the development of even more informative and easy-to-use visualization systems.

5 DANGER - A new Danish-German learner corpus

In connection with addressing our research questions based on empirical data, we also created a new DANish-GERman learner-corpus resource: DANGER. We are currently dealing with the possibility to host our corpus by the European research infrastructure initiative CLARIN (Hinrichs 2015). However, interested readers are also invited to approach us personally to get a free copy of the DANGER corpus for their own work. For this reason, we describe this resource in some more detail below.

5.1 Size and content

The video- and audio recordings of our participants constitute an informative and varied corpus that already comprises almost 5 hours of speech (4 hours 53 minutes in total) and consists of two major parts. The larger part (approx. 4 hours 20 minutes) consists of spontaneous L2 speech by Danish native speakers speaking (almost exclusively) German. The second smaller part (approx. 30 minutes) is read speech in the form of the target sentences that we analyzed for the present study.

As was described in 2.3, each pair of participants was asked to select any visualization technique, discuss it, and produce the two corresponding question and statement sentences using the prosody that they agreed upon in the discussion. After having read the two question and statement sentences, the participants evaluated the visualization technique, and then they proceeded to the next visualization technique. This procedure resulted in a lot of informal, highly interactive, and phenomenologically rich spontaneous speech surrounding the read target sentences. Although the interactions predominantly revolved around solving the tasks assigned to the speakers, some spontaneous speech sections also deal with other personal or study-oriented topics. Overall, the turn-taking in the spontaneous-speech sections is balanced, i.e. the two participants made, in all speaker pairs, similarly many and long contributions to the interaction. The speakers of each pair come from the same peer group; and in order to further facilitate their free, casual interaction, we chose surroundings that were familiar to the speakers and part of the their everyday lives, namely rooms around the campus of the University of Southern Denmark in Sønderborg.

5.2 Setup and recordings

The participants' interactions were recorded using both a digital voice recorder and a video camera. This means that both video and audio data can be examined independently as well as in time-aligned combination. Figure 8 above illustrates the setup of the recording sessions. As is shown, the recordings were made in silent, distraction-free surroundings. The participants sat side-by-side at a table about 1-2 meters away from the video camera. This distance was sufficient (a) for speakers to forget (after a short while) the presence of the video camera and (b) for the camera to capture all mimics, gestures, and upper-body movements of the participants. Only shifts in gaze direction cannot be consistently observed in our recordings as the participants' gaze is usually directed downwards toward the visualizations on the table

in front of them. The audio recordings are in stereo and available in the formats mp3 and wav; the videos were recorded in mp4 format.

5.3 Distinctive features of the corpus

The majority of learner corpora are collections of written speech in the form of essays, exams, and other forms of elicited language (see, for example, CLEG13 or FALKO for English and German, Walter et al. 2007). Furthermore, many corpora of written as well as spoken language focus on L2 English. Our combination of L1 Danish speakers producing read sentence and spontaneous dialogues in L2 German is probably fairly rare and thus particularly valuable. The LeaP corpus, a corpus of spoken language, comprises four speech styles, i.e. read speech, prepared speech, free speech from interviews, and nonsense word lists (see Gut 2009). In addition to including L2 German (though not from Danish native speakers), the LeaP corpus and our DANGER corpus share some common ground, for example, concerning the questions what aspects of prosody should be taught to L2 learners and how. However, the LeaP corpus and our corpus use different means and approaches to reaching these goals, from different target/age groups to different forms of data (audio vs. audio-visual). Another distinctive feature of our DANGER corpus is the use of the KIM model to annotate the target sentences prosodically at a fine level of detail.

So far, all 216 read-speech target sentences of our corpus have been transcribed, segmented, and fully annotated prosodically. A next step will be to provide at least a basic transcript of the remaining spontaneous-speech parts. These transcripts can then also be used for sound segmentation based on forced-alignment scripts like WebMaus (Kisler et al. 2012). Since our recordings include both audio and video recordings, the latter will allow the inclusion not only of verbal but also of non-verbal behavior into the transcripts. In this way, our growing corpus will serve as a source of data for various fields of study. Such fields include, but are not limited to, the study of embodiment, embodied metaphors, of gestures and mimics in general as well as of partner-orientation.

6 Future work

Future work will have to show to what extent some previous training can help students work themselves into a better mastery of the different visualization techniques over time. In this context, it would also be good to know which kinds of training are most effective. However, the main goal of our line of research remains to find visualizations that are not just consistently applicable but also intuitively understandable and hence achieve a high level of correct prosodic renderings without requiring substantial extra efforts from teachers and learners. Therefore, our main focus in future work will be to explore *additional* methods for visualizing intonation and other aspects of prosody and possibly develop new techniques, starting off from our findings regarding the superiority of the stylized "hat pattern" and its spatial proximity to the corresponding text.

For example, the MOMEL-based method of Hirst (2016) is similar to the "tadpole" visualization but intonationally less explicit. In view of our assumption that the "tadpole" system has overwhelmed the students with too much and too detailed information, Hirst's technique seems to be a good compromise. Although it still lacks the spatial proximity to the text that characterized the "hat pattern", it could perform similarly well as the latter. Hirst's technique is a very recent development. Future studies should also include some very old visualization techniques whose design is shaped by musical notations. It is possible that participants who can read music (and, due to music education at school, this a considerable proportion of potential users in the Western world) perform particularly well with such techniques, and perhaps even with the historically early model suggested by Steele (1775).

Moreover, we will also test to what extent the separation and/or subsequent presentation of prominence and intonation will help improve the consistency and reduce the errors of L2 learners in prosody production. In this context, we will include ideas about how stress-level information can be presented in the form of rhythmic patterns (cf. Wang et al. 2016).

In order to have a better starting point for tackling these issues, we currently conduct a follow-up study in which we supplement the KIM-based auditory-phonological correctness and consistency analyses with quantitative acoustic analyses. That is, we analyze the reference productions in terms of pitch ranges, movement velocities, durations, and acoustic-energy levels of words. Then, we relate these reference values to those measured in the corresponding words of our 18 participants. In this way, we will get a much more fine-grained picture about how the compared notation systems actually perform. What we found already is that phonological differences annotated on an auditory basis with the PROLAB labels of KIM consistently translate into significant differences in the acoustic domain. This

clearly indicates that the annotations for the present study were systematic and interpretable and thus created valid and reliable results.

Nevertheless, in order to increase the generalization of our results, another important task of follow-up studies will be to test the visualization techniques in combination with different target sentences. We will design our own target sentences that keep the advantages of the Berlin Sentences (e.g., a one-sentence-one-phrase make-up, common words, balanced pronounceability and comprehensibility) while avoiding disadvantages like variable sentence length, a variable number of post-nuclear unaccented syllables, and the use of different question types across visualization techniques. As was stated in 2.2.1, we will also represent each visualization technique by multiple questions and statements, thus further minimizing potential artifacts and biases that individual sentences can create in the prosodic patterns of sentences independently of their visualization. This will be done based on a reduced set of visualization techniques. In view of the present results, such a reduced set will probably consist of the "meandering text", the "hat pattern", and either "tadpoles" or a music-based prosody notation like that of Steele (1775).

Finally, it is a noticeable finding of the present study that all visualization techniques achieved consistency scores higher than 50 %. It is possible that this relatively good performance of all visualization techniques is supported by the fact that our participants shared the same native language, i.e. Danish. Future studies (with the same or different visualization techniques) will extend our L2 learner corpus with more data –of the kind collected in the present study (Danish native speakers speaking German) as well as data concerned with other L1 and L2 speakers. The latter data enables us to scrutinize the stabilizing effect of a common source language and address the cross-language generalization of our findings.

Acknowledgements

This study was conducted within the project *Improving Second Language Pedagogy at the Prosody-Pragmatic Interface by Using Human-Robot Interaction* (PI Kerstin Fischer). The project is financed by the Danish Council for Independent Research.

References

Abercrombie, D. 1991. *Fifty years in phonetics*. Edinburgh: Edinburgh University Press.

Anderson, K. 2004. "Can German intonation be taught?", in: Annas, R. (Ed.), *Deutsch als Herausforderung: Fremdsprachenunterricht und Literatur in Forschung und Lehre* (pp. 89-95). Stellenbosch: Sun.

Armstrong, L. & Ward, I.C. 1926. *Handbook of English Intonation*. Cambridge: Heffner.

Batliner, A. 1991. "Ein einfaches Modell der Frageintonation und seine Folgen". In: E. Klein, F. P. Duteil & K. H. Wagner (Eds.). *Betriebslinguistik und Linguistikbetrieb.* Tübingen: Niemeyer, 147-160.

Baumann, S., Grice M. & Steindamm, S. 2006. "Prosodic Marking of Focus Domains – Categorical or Gradient?", *Proc. 3rd International Conference of Speech Prosody, Dresden, Germany* 301–304.

Baumann, S., Niebuhr, O. & Schroeter, B. 2016. "Acoustic Cues to Perceived Prominence Levels – Evidence from German Spontaneous Speech", *Proc. 8th International Conference of Speech Prosody, Boston, USA* 1-5.

Bolinger, D. L. 1958. "A theory of pitch accent in English", *Word 14*, 109-149..

Beckman, M.E. & Pierrehumbert, J. 1986. "Intonational structure in English and Japanese", *Phonology Yearbook III*, 15-70.

Boren, M. & Ramey, J. 2000. "Thinking aloud: Reconciling theory and practice", *IEEE Transactions on Professional Communication* 43 (3), 261-78.

Breen, M., Dilley, L. C., Kraemer, J., & Gibson, E. 2012. "Inter-transcriber reliability for two systems of prosodic annotation: ToBI (Tones and Break Indices) and RaP (Rhythm and Pitch)", Corpus Linguistics and Linguistic Theory 8 (2), 277-312.

Caspers, J. & van Heuven, V. 1993. "Effects of time pressure on the phonetic realization of the Dutch accent-lending pitch rise and fall", *Phonetica* 50, 161-171.

Crystal, D. 1995. *The Cambridge Encyclopedia of Language*. Frankfurt: Campus.

Delattre, P. 1966. "Les Dix Intonations de base du français", *The French Review* 40, 1-14.

Essen, Otto von. 1956. *Grundzüge der hochdeutschen Satzintonation*. Ratingen: Henn.

Fagyal, Z. 1997. "Chanting Intonation in French", *University of Pennsylvania Working Papers in Linguistics* 4, 77-90.

Feldes, S. & Herzog, M. 1997. "Kategorisierung des Lautkontextes von deutschen Allophonen für die phonembasierte Erkennung", *Fortschritte der Akustik* 23, 551-552.

Fónagy, I. & Magdics, K. 1963. "Emotional patterns in intonation and music", *Zeitschrift für Phonetik und Allgemeine Sprachwissenschaft* 16, 293-326.

Fujimori, A., Yoshimura, N. & Yamane, N. 2015. “The Development of Visual CALL Materials for Learning L2 English Prosody”, *ICT for Language Learning. Florence, Italy*.

Gorjian, B., Hayati, A. & Pourkhoni, P. 2013. “Using Praat Software in Teaching Prosodic Features to EFL Learners”, *Procedia - Social and Behavioral Sciences* 84, 34-40.

Grabe, E. 1998. "Pitch accent realizations in English and German", *Journal of Phonetics* 26, 129–143.

Griesbach, H. 2000. *Bauplan Deutsch: Übungsgrammatik und Satzbauhelfer*. Frankfurt: Libri.

Grice, M., Baumann, S. & Benzmüller, R. 2005. “German Intonation in Autosegmental-Metrical Phonology”. In Jun, S.-A. (Ed.), *Prosodic Typology: The Phonology of Intonation and Phrasing*. Oxford University Press, 55-83.

Gut, U. & Bayerl, P.S. 2004. "Measuring the reliability of manual annotations of speech corpora", *Proc. 2nd International Conference of Speech Prosody, Nara, Japan*, 565-568.

Gut, U. 2009. *Non-native speech. A corpus-based analysis of phonological and phonetic properties of L2 English and German*. Frankfurt: Peter Lang.

Harst, E., Kaufmann, S., Moritz, U., Rodi, M., Rohrmann, L., Scherling, T. & Sonntag, R. 2015. *Linie 1 A1 - Deutsch in Alltag und Beruf. Kurs- und Übungsbuch mit DVD-ROM*. München: Klett.

Haugen, E., Joos, M. 1952. "Tone and intonation in East Norwegian", *Acta Philologica Scandinavica* 22, 41-64.Hirst, D. 2016. "On the automatic comparison and cloning of native and non-native speech prosody", *Proc 8th International Conference of Speech Prosody, Boston, USA*, 213-217.

Hinrichs, E. 2015. "CLARIN as European Adopter of RDA Outputs", *URL: https://www.rd-alliance.org/sites/default/files/attachment/07-CLARIN_ErhardHinrichs.pdf*

Isačenko, A. & Schädlich, H.-J. 1970. *A Model Of Standard German Intonation.* The Hague: Mouton.

Johner, Ch., Janke, M., Wand, M. & Schultz, T. 2012. “Inferring Prosody from Facial Cues for EMG-based Synthesis of Silent Speech”, *Proc. 4th International Conference on Applied Human Factors and Ergonomics, Las Vegas, USA* 1-10.

Jones, D. 1909. *Intonation curves, a collection of phonetic texts, in which intonation is marked throughout by means of curved lines on a musical stave*. Leipzig and Berlin: B. G. Teubner.

Kisler, T., Schiel, F. & Sloetjes, H. 2012. "Signal processing via web services: the use case WebMAUS", *Proc. Digital Humanities, Hamburg, Germany*, 30-34.

Klinghardt, H. 1923. *French Intonation Exercises*. Cambridge: Heffer.

Klinghardt, H. 1927. *Übungen im Deutschen Tonfall*. Leipzig: Meyer.

Kohler, K. J. 1990. “Macro and micro F0 in the synthesis of intonation”. In J. Kingston & M. E. Beckamn (Eds.), *Papers in Laboratory Phonology I*. Cambridge: CUP, 115-138.

Kohler, K.J. 1991. "The interaction of fundamental frequency and intensity in the perception of intonation", *Proceedings 12th ICPhS, Aix-en-Provence, France*, 186-189.

Kohler, K. J. 1997. “Modelling prosody in spontaneous speech”. In Y. Sagisaka, N. Campbell & N. Higuchi (Eds.), *Computing Prosody. Computational models for processing spontaneous speech*. New York: Springer, 187-210.

Kügler, F. 2008. “The role of duration as a phonetic correlate of focus”, *Proc. 4th International Conference of Speech Prosod, Campinas, Brazil* 591-594.

Kügler, F., Smolibocki, B., Baumann, S., Braun, B., Grice, M., Jannedy, S., Niebuhr, O., Peters, J., Schweitzer, K., Schweitzer, A. & Wagner, P. 2015. “DIMA - Annotation guidelines for German intonation”, *Proc. 17th ICPhS, Glasgow, Scotland*.

Ladd, R. D. 1978. “Stylized intonation”, *Language* 54, 517-540.

Ladd, R. D. 2008: *Intonational Phonology*. Cambridge: Cambridge University Press.

Mehlhorn, G. & Trouvain, J. 2007. “Sensibilisierung von Lernenden für fremdsprachliche Prosodie”, *Zeitschrift für Interkulturellen Fremdsprachenunterricht* 12, 1-15.

Möbius, B. 1993. *Ein quantitatives Modell der deutschen Intonation: Analyse und Synthese von Grundfrequenzverläufen.* Tübingen: Niemeyer.

Niebuhr, O. 2010. “On the phonetics of intensifying emphasis in German”, *Phonetica* 67, 170-198.

Niebuhr, O. 2013. “The acoustic complexity of intonation”. In E.-L. Asu, P. Lippus (Eds.), *Nordic Prosody XI*. Frankfurt: Peter Lang, 15-29.

Niebuhr, O. 2015. “Gender Differences in the Prosody of German Questions”. *Proc. 18th International Congress of Phonetic Sciences, Glasgow, UK* 1-5.

Niebuhr, O. & Dilley, L. 2016. "Prosody and Perception - Approaching the Real Complexity of Pitch-Accent Perception", *Paper presented at the Aix Summer School on Prosody: Methods in Prosody and Intonation Research*. URL: http://aixprosody2016.weebly.com/uploads/2/6/4/4/26448693/abstract_dilleyniebuhr_perceptioni_ii.pdf

O’Connor, J. D. & Arnold, G. 1973. *Intonation of Colloquial English.* London: Longman.

Peters, B. 1999. “Prototypische Intonationsmuster in deutscher Lese- und Spontansprache”, *Arbeitsberichte des Instituts für Phonetik und Digitale Sprachverarbeitung der Universität Kiel (*AIPUK) 34, *1-177*.

Peters, B., & Kohler, K. J. 2004. “Trainingsmaterialien zur prosodischen Etikettierung mit dem Kieler Intonationsmodell KIM”. URL: www.ipds.uni-kiel.de/kjk/pub_exx/bpkk2004_1/TrainerA4.pdf

Pierrehumbert, J. 1980. *The phonology and phonetics of English intonation*. PhD thesis, MIT.

Pike, K. 1945. *The Intonation of American English*. Michigan: University of Michigan Press.

Prom-om, S. & Xu, Y. 2010. “The qTA Toolkit for Prosody: Learning Underlying Parameters of Communicative Functions Through Modeling”, *Proc. 5th International Conference of Speech Prosody, Chicago, USA* 1-5.

Rathcke, T.V. 2013. "On the Neutralizing Status of Truncation in Intonation: A Perception Study of Boundary Tones in German and Russian", *Journal of Phonetics* 41, 172-185.

Selting, M., Barth-Weingarten, D., Bergmann, J., Bergmann, P., Birkner, K., Couper-Kuhlen, E., Deppermann, A., Gilles, P., Günthner, S., Hartung, M., Kern, F., Mertzlufft, Ch., Meyer, Ch., Morek, M., Oberzaucher, F., Peters, J., Quasthoff, U., Schütte, W., Stukenbrock, A. & Uhmann, S. 2009. “Gesprächsanalytisches Transkriptionssystem 2 (GAT 2)”, *Gesprächsforschung - Online-Zeitschrift zur verbalen Interaktion* 10, 353-402.

Sotschek, J. 1984. “Sätze für Sprachgütemessung und ihre phonologische Anpassung an die Deutsche Sprache”, *Tagungsband DAGA: Fortschritte der Akustik* 873-876.

Steele, J. 1775/1969. *An essay towards establishing the melody and measure of speech*. Menston: The Scholar Press.

‘t Hart, J., Collier, R. & Cohen, A. 1990. *A perceptual study of intonation: an experimental phonetic approach to speech melody*. Cambridge: Cambridge University Press.

Wahlster, W. (Ed.). 2000: *Verbmobil: Foundations of Speech-to-Speech Translation.* Berlin et al.: Springer.

Walter, M., Schmidt, K., Lüdeling, A., Byrnes, H. & Maxim, H. 2007. "Falko-Georgetown-Longitudinalkorpus", *URL: https://www.linguistik.hu-berlin.de/de/institut/professuren/korpuslinguistik/forschung/falko/FalkoGeorgetownDokumentation.pdf*

Wang, H., Mok, P., & Meng, H. 2016. "Capitalizing on musical rhythm for prosodic training in computer-aided language learning", *Computer Speech & Language 37*, 67-81.

Wightman, C. W., Shattuck-Hufnagel, S., Ostendorf, M. & Price, P. J. 1992. “Segmental durations in the vicinity of prosodic phrase boundaries”, *Journal of the Acoustical Society of America* 91, 1707–1717.